Evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal

a case study from health science

Madsen, Heidi Holst; Madsen, Dicte; Gauffriau, Marianne

Published in:
F1000Research

DOI:
10.12688/f1000research.8913.2

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

RESEARCH ARTICLE

REVISED Evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal – a case study from health science [version 2; referees: 2 approved, 1 approved with reservations]

Previously titled: Evaluation of unique identifiers used for citation linking

Heidi Holst Madsen[1], Dicte Madsen[1,2], Marianne Gauffriau[1,3]

[1]Faculty Library of Natural and Health Sciences, Copenhagen University Library, The Royal Library, Copenhagen, DK-2200, Denmark
[2]Copenhagen Business School Library, Frederiksberg, DK-2000, Denmark
[3]SUND Research & Innovation, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, DK-2200, Denmark

v2 **First published:** 29 Jun 2016, **5**:1539 (doi: 10.12688/f1000research.8913.1)
Latest published: 06 Sep 2016, **5**:1539 (doi: 10.12688/f1000research.8913.2)

Abstract

Unique identifiers (UID) are seen as an effective key to match identical publications across databases or identify duplicates in a database. The objective of the present study is to investigate how well UIDs work as match keys in the integration between Pure and SciVal, based on a case with publications from the health sciences. We evaluate the matching process based on information about coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match keys. We analyze this information to detect errors, if any, in the matching process. As an example we also briefly discuss how publication sets formed by using UIDs as the match keys may affect the bibliometric indicators number of publications, number of citations, and the average number of citations per publication.

The objective is addressed in a literature review and a case study. The literature review shows that only a few studies evaluate how well UIDs work as a match key. From the literature we identify four error types: Duplicate digital object identifiers (DOI), incorrect DOIs in reference lists and databases, DOIs not registered by the database where a bibliometric analysis is performed, and erroneous optical or special character recognition.

The case study explores the use of UIDs in the integration between the databases Pure and SciVal. Specifically journal publications in English are matched between the two databases. We find all error types except erroneous optical or special character recognition in our publication sets. In particular the duplicate DOIs constitute a problem for the calculation of bibliometric indicators as both keeping the duplicates to improve the reliability of citation counts and deleting them to improve the reliability of publication counts will distort the calculation of average number of citations per publication.

The use of UIDs as a match key in citation linking is implemented in many settings, and the availability of UIDs may become critical for the inclusion of a publication or a database in a bibliometric analysis.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 06 Sep 2016		✓ report	
version 1 published 29 Jun 2016	? report	? report	✓ report

1 **Marion Schmidt**, German Centre for Higher Education Research and Science Studies (DZHW) Germany

2 **Keith G. Jeffery**, Keith G. Jeffery Consultants UK

3 **Sarah L. Shreeves**, University of Miami USA

Discuss this article

Comments (0)

Corresponding author: Marianne Gauffriau (marianne.gauffriau@sund.ku.dk)

How to cite this article: Madsen HH, Madsen D and Gauffriau M. **Evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal – a case study from health science [version 2; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2016, **5**:1539 (doi: 10.12688/f1000research.8913.2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 29 Jun 2016, **5**:1539 (doi: 10.12688/f1000research.8913.1)

REVISED **Amendments from Version 1**

The reviews from Marion Schmidt and Keith G. Jeffery have raised some issues in version 1. In version 2 the major changes are:

The term *citation linking* has caused confusion as citation can be understood as the reference from one paper (citing document) to another paper (cited document). By citation linking we have meant the linking of representations of identical publications across databases. We now have minimized the use of the term citation linking. Instead we use match key and matching process.

The objective was formulated somewhat broadly and could give the impression that the study included a thorough analysis of the coverage of SciVal/Scopus and the study yielded results which could be generalized to other research fields. We have specified the objective to clearly show that we evaluate how well UIDs work as match key. Furthermore, we do only work with one case – the integration between Pure and SciVal for a publication set from health sciences. As a consequence of the reformulated objective, we have changed the title and abstract accordingly.

The description of a publication and its UIDs is elaborated. We did not analyze the full content of a publication but only parts of its metadata representations in Pure and SciVal.

The literature review shows the big picture and now also includes more details on the studies included. None of the studies in the review were directly comparable to our study as we analyzed how well UIDs work as match key in the matching process between Pure and SciVal.

Please see the response to the reviewers for a detailed record of differences between version 1 and 2.

See referee reports

Introduction

Unique identifiers (UIDs) have been introduced for more and more entities, e.g. Open Researcher and Contributor ID (ORCID) for researchers, and digital object identifiers (DOI) for research publications, etc. One advantage of UIDs is that integrations between databases potentially can be done much more efficiently. This is stressed in a recent evaluation of metrics in research evaluations (Wilsdon *et al.*, 2015) p. 15–22, 145).

The purpose of the present study is to find out how well UIDs work as match keys for publications and thus to create publication sets for bibliometric analysis. Traditionally, this is done via a match key based on bibliographic information such as author, title, etc. The exact method is rarely described. An exception is the evaluation of the Danish Council for Independent Research (Schneider *et al.*, 2014, p. 36–38).

UIDs are simple match keys compared to the traditional method (e.g. Olensky *et al.*, 2015). We explore how the method works in the integration between the current research information system (CRIS), Pure, and the bibliometric research evaluation tool, SciVal, (Elsevier, 2014). SciVal builds on data from the citation index Scopus, and Pure provides a uniform identification of researchers and the organizational structure at a university. UIDs make it easy to export a publication set from Pure to SciVal for bibliometric analysis (Figure 1). An alternative is to define the publication set, e.g. the publications from a department, in Scopus or Web of Science (WoS). This is often a resource-demanding task as researchers do not always register their affiliations correctly and consistently in publications (e.g. Moed *et al.*, 1995, p. 390).

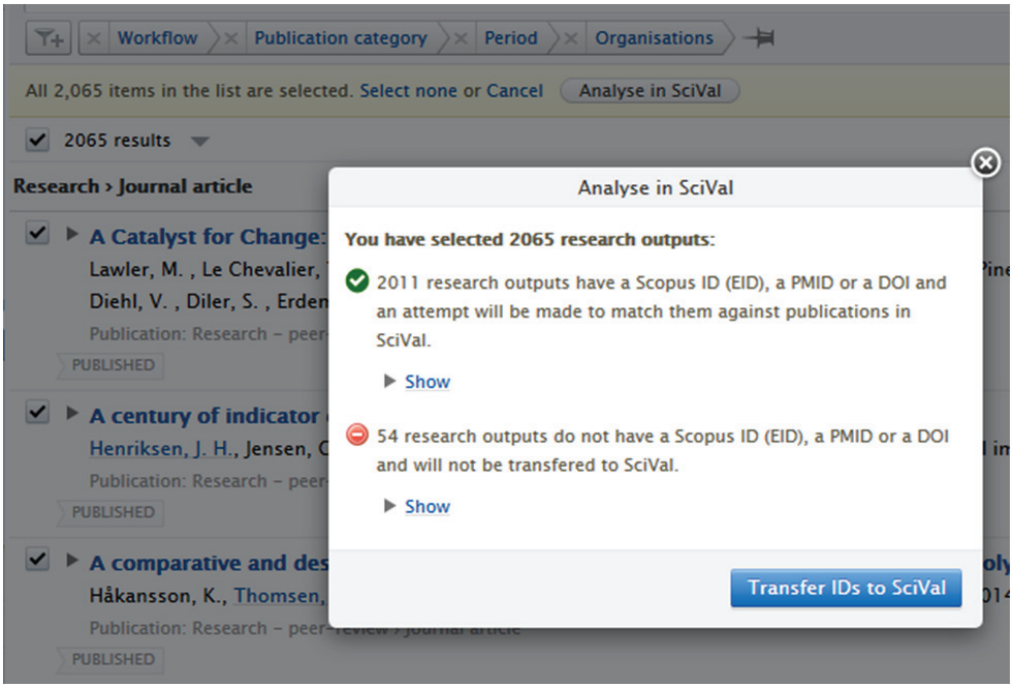

Figure 1. Screenshot from Pure. Publication set automatically analyzed for UIDs before export to SciVal, August 2015. Source: Pure from Elsevier, version 4.23.1, local installation at University of Copenhagen.

A widely used UID for research publications is the DOI. It was launched in 2000 (International DOI Foundation, 2015, sec. 1.2) and is by now assigned to publications by more than 5,000 publishers from the big players, e.g. Elsevier, to small societies, e.g. Danish Chemical Society (CrossRef, 2015b). Other UIDs for research publications such as arXiv ID from 1991 (arXiv.org, 2015), PubMed ID (PMID) from 1997 (Ostell *et al.*, 1998, p. 27), and Scopus ID (EID) from 2004 (Elsevier, 2004) are not as prevalent as the DOI. The UIDs are not assigned to the publication itself (the content), but to the metadata of a publication. How and if this metadata is recorded can differ between databases.

The integration between Pure and SciVal was launched in October 2014 and DOI, PMID and EID were used as match keys. From October 2015 Elsevier matches publications in Pure with Scopus and attributes EIDs to publications in Pure. And from March 2016 the integration between Pure and SciVal is based on EID alone. The latter will not affect the present study as we analyze publication sets downloaded in August and December 2015. It means that our study is a documentation of the functionality of the first matching processes between Pure and SciVal and our results can serve as background for new evaluations. Is the new matching process based on EIDs better? We do not know. EIDs from Scopus are added to publications in Pure and these publications can subsequently be matched with SciVal which builds on Scopus. It would be interesting to see if this matching process solves some of the problems of having UIDs as match keys which we show in this study or if it causes new problems.

Objective

The objective of the present study is to investigate how well UIDs work as match keys in the integration between Pure and SciVal, based on a case with publications from the health sciences. We evaluate the matching process based on information about coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match keys. We analyze this information to detect errors, if any, in the matching process.

The matching of publications across databases can be used to create publication sets for bibliometric analysis. As an example we also briefly discuss how publication sets formed by using UIDs as the match keys may affect the bibliometric indicators number of publications, number of citations, and the average number of citations per publication.

Limitations

We analyzed a case study from the health sciences and our results should be interpreted in context of the case. The results cannot be generalized, e.g. to other research areas. Furthermore we analyzed information on coverage, precision, and characteristics to detect errors in the matching process between Pure and SciVal. This type of information did not allow us to detect all types of errors. If a publication was assigned a wrong UID in one of the databases and thus wrongly matched we may not have detected this. In a sample of the apparently correctly matched publications we checked if the matched publications were identical and found no errors.

Throughout the study we use the term publication even though we did not analyze the full content of a publication but only parts of its metadata representations in Pure and SciVal; mainly UIDs, publication type, language, and journal. If a publication had more UIDs assigned to it in one of the databases we assumed that each of the UIDs represented the same publication. We do not know the details of the matching process between Pure and SciVal, but assumed that if one UID did lead to a match then the other UIDs assigned to the same publication were not taken into account. From SciVal we did not get any reports on contradicting UIDs for a publication. Such contradictions could question if there was a match or not.

We describe the characteristics of two publication sets from Pure and the two matched publication sets from SciVal, but it is beyond the scope of this paper to do a thorough analysis of the metadata quality of the Pure and SciVal publication sets. Also we limit the discussion of bibliometric indicators to three basic bibliometric indicators. More advanced indicators and their construction in SciVal are not discussed.

Methods

The objective was addressed in a literature review and a case study. The literature review gave us an indication of the use of UIDs as keys for matching identical publications in one or more databases, and an overview of the precision of the method. As the integration between Pure and SciVal is relatively new and evaluations are not yet reported in the literature, we conducted our own case study to see if the implementation of UIDs as match keys between SciVal and Pure confirms what other studies have found.

Literature review

Evaluations of UIDs as match keys to detect identical publications in different databases or duplicates in one database were identified. Information on search terms, search strategy and databases are given below. UIDs as match keys have been used for many applications, but our focus was on research publications, with a particular interest in how the method may affect bibliometric analysis. Thus, the search was limited to studies where UIDs of publications are used as the match keys or part of the match key, and in which the method is analyzed and discussed in some detail.

An exploratory search showed that the terminology for applying match keys to detect identical publications is not consistent. It is called citation linking or reference linking or simply matching or linking. Matching within the same database is called deduplication. The term citation matching is also used, but often for the more specific purpose where citing and cited publications are matched. We also saw examples of more general terminology, namely integration or interoperability between databases or retrieval strategy. In our subsequent searches the different terms for applying match keys to identify identical publications were combined (Boolean AND) with different terms for UID: unique, identifier, DOI, PMID. This gave us an idea of which databases use UIDs as match keys, e.g. CrossRef, Mendeley, and Altmetric.com. We also included these databases as search terms and combined them with the different terms for UIDs.

The searches were conducted in WoS (https://login.webofknowledge.com/), Scopus (https://www.scopus.com/), and Google Scholar (https://scholar.google.dk/). No range of years was specified. If no relevant publications were found in WoS and Scopus, we continued the search in Google Scholar. This means that not only peer-reviewed research but also a few reports were included in the literature review. In relevant publications, we manually scanned references and citations for other relevant publications. The searches were done in August and September 2015 followed by later supplementary searches based on the references found in August and September.

Case study

In the case study we explored information about coverage, precision, and characteristics of publications matched versus not matched in the integration between Pure and SciVal to detect errors, if any, in the matching process. Our publication set is from the Department of Clinical Medicine (DoCM) at University of Copenhagen (UCPH). DoCM registers approximately 2,000 research publications in the UCPH Pure database per year. The majority are peer-reviewed journal publications in English. As this type of publication and the health sciences are well-covered in Scopus/SciVal (Mongeon & Paul-Hus, 2016, p. 218–219+222; Valderrama-Zurián *et al.*, 2015, p. 570–571), we expected the DoCM publication set to be well-fitted for our purpose, namely to explore the matching process, rather than how well Scopus/SciVal covers publications from a department.

The publication set was limited to research publications published in 2014, registered and validated in Pure. Publications published before 2014 were not included as these have been validated at department or group level and the data quality is not consistent as no common practice was in place. The validation of publications from 2014 was undertaken by the authors of this article and three information specialists from the University Library as a service for the Faculty of Health and Medical Sciences. As part of the validation process, language and publication type was determined according to the categories available in Pure. This information is utilized in the Results section. However, the focus of the validation was not UIDs as match keys, and fields for UIDs were not mandatory. If PMID or EID was registered in Pure, it is most likely because the publication was imported from PubMed or Scopus. A publication in Pure without a UID may not have a UID, or the UID is simply not registered in Pure. As mentioned in the Introduction, from October 2015 Elsevier matches publications in Pure with Scopus and attributes EIDs to new publications and retrospectively.

Our choice of case implies some limitations. The publication sets have too few non-journal publications to draw conclusions on their coverage and the precision in publications matched versus not matched with UIDs as match keys. Furthermore, the publication year 2014 gives the publications too short of a time since publication to obtain robust citation counts.

The case study alone did not lead to generalizable results, but the results were compared to findings from the literature review to identify trends and compatibility with previous studies.

Before we analyzed the outcome of the matching process based on UIDs as match keys, we downloaded, merged, and cleaned data from Pure and SciVal. This process was carried out in August 2015 (n=2068) and repeated in December 2015 (n=2066). It is possible for researchers and administrative staff to make retrospective changes to the registrations in Pure; this is the most plausible explanation for the lower number of publications in December.

Data software

- Pure local installation at University of Copenhagen, version 4.22.1 for the August download and version 4.23.1 for the December download (data download)
- SciVal June 8, 2015, and September 30, 2015 releases (data analysis and download)
- Microsoft Excel 2007 (data cleaning and analysis)

Data download, merging and cleaning

Raw data was downloaded from Pure in August and December 2015 using the following filters:

- Organisational unit = Department of Clinical Medicine
- Publication category = Research
- Publication statuses and dates > Latest > Date: Selected range = 2014
- Workflow = Validated

To fit relevant data in just one worksheet in Excel and be able to create a .csv file, most of the data columns were deleted, and only the following kept:

- Access to electronic version (full text) > DOI (Digital Object Identifier)-0
- journalAssociation.title
- pages
- persons[0].lastName
- typeClassification.typeClassification
- title
- id [=Pure ID]
- Source[sourceId]: PubMed [=PMID]
- Source[sourceId]: Scopus [=EID]
- language.language

Due to an error in the Copenhagen University Pure at the time, it was not possible to download a full data report of publications with the DOI column. Instead, first an ungrouped raw data report was downloaded, then the same report grouped on DOI. The two reports were matched on Pure ID to create one list with DOI data where available.

The Data set 1 DoCM Pure data August.csv and Data set 2 DoCM Pure data December.csv files comprise our "raw" Pure data – ever

so slightly tidied to a) create one full data report with DOI where available, b) fit relevant columns in one worksheet to be able to create a .csv file.

The Pure "raw" data was furthermore cleaned by:

- Removing superfluous spaces at the end of DOIs to be able to match DOIs in the Pure data with the DOIs in the SciVal data.

After the Pure data was sent to SciVal for analysis, the resulting SciVal publication sets (August and December) were downloaded from SciVal with the following information:

- Title
- Authors
- Journal title
- Citations
- Pages
- DOI
- Publication-type
- EID [=Scopus ID]
- PubMed ID [=PMID]

The Data set 3 DoCM SciVal data August.csv and Data set 4 DoCM SciVal data December.csv files comprise our raw SciVal data.

The SciVal raw data was furthermore cleaned by:

- Removing "2-s2.0-" from the EIDs to be able to match with the EIDs in the Pure data.
- Duplicate DOIs were identified to remove superfluous/ irrelevant publications:
    - Article vs. Article in Press (Article kept in data set)
    - Publication duplicates (if one duplicate had a PMID, that is the one we kept; otherwise we randomly selected which duplicate to keep)
    - Publication vs. publication attributed wrong ID in Scopus/SciVal and not occurring in the Pure data set (Publication in Pure data set kept)
    - Publication registered as one publication type vs. same publication registered as another publication type (duplicate with same publication type as in the Pure data set was kept)
    - Author's reply (not in Pure data set) having same DOI as the publication (in Pure data set) it relates to (publication in Pure data set kept).

A note on some Article in Press occurrences in the SciVal data:

1. Sometimes SciVal imports only the Article in Press instance of an article in Scopus (instead of the published article instance), or the article is registered in Scopus only as Article in Press, although it is published.
2. During an automatic update in June 2015 of the UCPH Pure, a number of validated publications were changed from published in 2015 to published in 2014, although really they were published in 2015. As such, they should not have been part of our Pure publication set to begin with.

In the Results section, characteristics for three groups of publications are shown: Publications with UID exported from Pure to SciVal and matched, publications with UID exported from Pure to SciVal and not matched, and publications without UID not exported from Pure to SciVal. The publications without UID (DOI, PMID, or EID) were extracted from the cleaned Data set 1 and 2 with Pure data. To identify publications exported from Pure to SciVal and matched, we compared UIDs (DOI, PMID, or EID) in the cleaned Data set 3 with UIDs in the cleaned Data set 1 (August download). Publications in Data set 1 with no corresponding UID in Data set 3 constitute publications with UID exported from Pure to SciVal and not matched. This was repeated for Data set 2 and 4 (December download). The EIDs attributed automatically to publications in Pure were not visible in our raw data. We found 32 publications in Data set 4 from SciVal which must have an EID in Pure and SciVal as no other UID was assigned to them. Finally we compared UIDs in the cleaned Data set 3 and 4 to identify publications matched in SciVal in December but not in August.

Results

Results of the literature review

The literature review shows two trends. Firstly, the publication year of relevant studies is 2011 or later. Older UIDs such as arXiv ID and PMID do not seem to have the same momentum as DOI. Secondly, the use of UIDs as match keys in bibliometric studies and citation indexes seems under-reported. A possible explanation is that the commercial players do not publish their methodologies in full detail (Olensky, 2014, p. 3). However, in a study from 2015, two bibliometric research groups provide documentation for how they use DOI as part of their match keys (Olensky *et al.*, 2015, p. 7–9).

If we do not focus on the matching process in citation indexes and bibliometric analysis alone, we find an increasing number of tools for handling and analyzing research publications, e.g. CrossRef's cited-by links (CrossRef, 2015a) and Altmetric.com's embeddable badges (Altmetric.com, n.d.). Evaluations of these databases were also included in our literature review. But also for these tools evaluations of UIDs as match keys are rare.

We included 17 studies in the literature review. Two of the studies' main focus was to evaluate the integration of UIDs in existing systems and are in that sense comparable to our objective. For a small publication set, Zahedi *et al.* (2014) evaluated the quality of Mendeley metadata by a comparison with Web of Science metadata. The DOIs did not match for 15 publications (8%). Missing DOI or erroneous special character recognition were identified as error types. Franceschini *et al.* (2015) showed that a DOI in Scopus can point to multiple not identical publications. These error types were included in the overview in Table 1.

The rest of the studies used UIDs as match keys for different purposes and in doing so discussed how well the matching process worked. Different from our study, the main focus regarding

the UIDs was the coverage. For all the studies, UIDs did not cover all publications potentially relevant for the analyses. Some studies only included publications with UIDs, and others supplemented the match keys with other types of bibliographic information to include publications without UIDs. We provide a short summary of how UIDs were used in the studies. Keep in mind that UIDs are used in many other studies and tools, but with very limited or no documentation and discussion of the UIDs.

The vast majority of the studies (ten) analyzed altmetric scores or altmetric methods. They used UIDs (DOI, PMID, and arXiv ID) sometimes in combination with other match key elements to match a publication set from WoS, Scopus, arXiv.org, a local database, or specific journals with identical publications in Mendeley, Almetric.com, Twitter, Wikipedia, Impact Story, CiteULike, BibSonomy, Connotea, or a selection of Elsevier databases (Bar-Ilan *et al.*, 2012; Costas *et al.*, 2014; Haunschild & Bornmann, 2016; Haustein *et al.*, 2014; Haustein & Siebenlist, 2011; HEFCE, 2015; Kraker *et al.*, 2015; Nuredini & Peters, 2015; Zahedi *et al.*, 2014a). Two studies analyzed deduplication via UIDs (DOI, PMID, and arXiv ID) and other match key elements in Mendeley (Hammerton *et al.*, 2012) and for a metasearch engine which covered five biomedical datadases (Jiang *et al.*, 2014). One study created a citation index and used DOIs and other match key elements to match citing and cited publications (Kim & Kim, 2013). One report tested the interoperability between Researchfish and local CRISs via DOI or PMID (Research Councils UK, 2015). Finally, two studies reported on missing citations in WoS and Scopus. DOI was used to match identical publications in the two databases (Franceschini *et al.*, 2013; Franceschini *et al.*, 2014). Nine studies, in addition to the coverage of UIDs, also addressed the precision or types of errors when UIDs are used as match keys. This information is central for our study. The types of errors are summarized in Table 1. As shown above the studies have used UIDs for different purposes. Still, we recognized the error types as relevant for our objective.

Results of the case study

In the case study we analyzed research publications (co-)authored by the Department of Clinical Medicine (DoCM) at the University of Copenhagen, published in 2014, and registered and validated in Pure. We evaluated the matching process based on information about coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match keys to detect errors, if any, in the matching process.

The share of publications matched between Pure and SciVal, or the coverage, is 85.6% in August and 89.3% in December.

There are precision issues for a minor part of the publication sets. Three of the error types reported by other studies (Table 1) are also present in our publication sets.

Duplicate DOIs (Table 1): An automatic report from SciVal states that 1837 publications (August) and 1876 publications (December) are matched with our Pure publication sets. These numbers are inflated due to DOI duplicates (Table 2).

Table 1. Precision - types of errors.

Error due to	Reported by
Duplicate DOIs	(Costas *et al.*, 2015, p. 2015) (Zahedi *et al.*, 2014a, p. 1495) (Haustein *et al.*, 2014, p. 1) (Franceschini *et al.*, 2015)
Incorrect DOIs in reference lists and databases	(Franceschini *et al.*, 2013, p. 2153) (Costas *et al.*, 2015, p. 2015) (Zahedi *et al.*, 2014a, p. 1495)
DOIs not registered by the database where a bibliometric analysis is performed	(Haustein & Siebenlist, 2011, p. 449) (Bar-Ilan *et al.*, 2012, p. 101) (Franceschini *et al.*, 2014, p. 759) (Zahedi *et al.*, 2014b)
Erroneous optical or special character recognition	(Haustein & Siebenlist, 2011, p. 449) (Zahedi *et al.*, 2014b)

Table 2. Number and type of DOI duplicates in SciVal publication sets.

	August	December
Matched unique publications	1770	1844
DOI duplicates in SciVal	67	32
Matched publications including duplicates	1837	1876
Types of duplicates		
- articles-in-press/published articles	50	20
- articles/articles	12	7
- articles/articles with wrong DOI in Scopus and not in our Pure publication set	5	2
- articles/same publications but recorded as another publication type	0	3
Total	67	32

From August to December the number of duplicates decreases partly due to Scopus's automatic cleaning process, where an Article in Press is deleted after the published version is registered in Scopus. We have discussed our results with consultants from Elsevier's SciVal team and this has led to a correction of some of the other duplicates. It may also have had an effect that Elsevier in October 2015 started adding EIDs automatically to publication records in Pure.

Incorrect DOIs in reference lists and databases & DOIs not registered by the database where a bibliometric analysis is performed (Table 1): In the August and December publication sets, respectively 5 and 2 of the DOI duplicates are examples of publications assigned a wrong DOI in Scopus (Table 2). For a 10% sample of the remaining matched publications in the August and December publication sets we verified the DOIs. The publications were sorted by DOI and every tenth publication was searched in Scopus, PubMed, and CrossRef where title, authors, journal, and start page were compared. No errors were identified, but two publications did not have their DOIs registered in Scopus. Furthermore, we checked the 77 publications not matched in SciVal in August but matched in December. Of these, 36 publications have a DOI in our Pure publication set. No errors were found in Scopus. But as the publications were unmatched in August, DOIs or other UIDs must have been missing or been incorrect in Scopus in August or the publications were not indexed in Scopus in August.

We now turn to the characteristics of the publications in our publication sets. In Table 3–Table 7, the general characteristics of publications matched versus not matched in the integration between Pure and SciVal are presented. We have a particular interest in the publications' UIDs as these are essential for a possible match. Publication type and language can give us an indication of whether all potential matches are made. We expected journal publications in English to be matched because they are well-covered in Scopus/SciVal. Table 3 gives an overview of how many publications were matched and unmatched. For the unmatched publications we also show how many have a UID.

In Table 4a & Table 4b we focus on the types of UIDs for the matched and the unmatched publications.

Table 3. Export from Pure to SciVal - number of matched publications, unmatched publications with UID, and unmatched publications without UID. Download from August and December 2015.

	August 2068 publications		December 2066 publications	
Matched	1770	86%	1844	89%
Unmatched with UID	244	12%	178	9%
Unmatched without UID	54	3%	44	2%

Table 4a. UID type for publications with UID exported from Pure to SciVal and matched. Download from August and December 2015.

	August 1770 publications		December 1844 publications	
DOI	1726	98%	1757	95%
PMID	1659	94%	1720	93%
EID	9	1%	-	-
Any UID	1770	100%	1844	100%

Table 4b. UID type for publications with UID exported from Pure to SciVal and not matched. Download from August and December 2015.

	August 244 publications		December 178 publications	
DOI	104	43%	71	40%
PMID	219	90%	155	87%
EID	1	0%	-	-
Any UID	244	100%	178	100%

Table 5a. Publication type in Pure of publications with UID exported from Pure to SciVal and matched. Download from August and December 2015.

	August 1770 publications		December 1844 publications	
Contribution to journal:	**1765**	**>99%**	**1836**	**>99%**
Journal article	1592	90%	1650	89%
Letter	25	1%	25	1%
Review	102	6%	114	6%
Editorial	11	1%	12	1%
Comment/debate	35	2%	35	2%
Conference abstract in journal	0	0%	0	0%
Book/anthology/thesis/report:	**0**	**0%**	**2**	**<1%**
Book	0	0%	1	<1%
Anthology	0	0%	0	0%
Report	0	0%	0	0%
Doctoral thesis	0	0%	1	<1%
Contribution to book/anthology/report:	**4**	**<1%**	**4**	**<1%**
Book chapter	2	<1%	2	<1%
Article in proceedings	2	<1%	2	<1%
Encyclopedia chapter	0	0%	0	0%
Contribution to conference:	**0**	**0%**	**1**	**<1%**
Poster	0	0%	0	0%
Conference abstract for conference	0	0%	0	0%
Paper	0	0%	1	<1%
Other	**1**	**<1%**	**1**	**<1%**

Table 5b. Publication type in Pure of publications with UID exported from Pure to SciVal and not matched. Download from August and December 2015.

	August 244 publications		December 178 publications	
Contribution to journal:	**237**	**97%**	**172**	**97%**
Journal article	212	87%	157	88%
Letter	2	1%	2	1%
Review	20	8%	9	5%
Editorial	2	1%	2	1%
Comment/debate	0	0%	0	0%
Conference abstract in journal	1	<1%	2	1%
Book/anthology/thesis/report:	**1**	**<1%**	**0**	**0%**
Book	1	<1%	0	0%
Anthology	0	0%	0	0%
Report	0	0%	0	0%
Doctoral thesis	0	0%	0	0%
Contribution to book/ anthology/report:	**6**	**2%**	**6**	**3%**
Book chapter	4	2%	4	2%
Article in proceedings	1	<1%	1	1%
Encyclopedia chapter	1	<1%	1	1%
Contribution to conference:	**0**	**0%**	**0**	**0%**
Poster	0	0%	0	0%
Conference abstract for conference	0	0%	0	0%
Paper	0	0%	0	0%
Other	0	0%	0	0%

Table 5c. Publication type in Pure of publications without UID not exported from Pure to SciVal. Download from August and December 2015.

	August 54 publications		December 44 publications	
Contribution to journal:	**33**	**61%**	**25**	**57%**
Journal article	22	41%	16	36%
Letter	2	4%	2	5%
Review	2	4%	1	2%
Editorial	1	2%	0	0%
Comment/debate	0	0%	0	0%
Conference abstract in journal	6	11%	6	14%
Book/anthology/thesis/report:	**7**	**13%**	**6**	**14%**
Book	0	0%	0	0%
Anthology	1	2%	1	2%
Report	1	2%	1	2%
Doctoral thesis	5	9%	4	9%
Contribution to book/ anthology/report:	**8**	**15%**	**8**	**18%**
Book chapter	6	11%	6	14%
Article in proceedings	2	4%	2	5%
Encyclopedia chapter	0	0%	0	0%
Contribution to conference:	**5**	**9%**	**4**	**9%**
Poster	3	6%	3	7%
Conference abstract for conference	1	2%	1	2%
Paper	1	2%	0	0%
Other	1	2%	1	2%

DOI is the most common UID (Table 4a) but nearly as many publications have a PMID. This was expected as the majority of the publications were imported from PubMed to Pure in our specific publication set. In the August publication set, very few publications had an EID; most likely because Scopus is not commonly used for import to Pure by DoCM. In the December set we could not analyze the EIDs, as automatically attributed EIDs are not shown in our Pure reports of raw data. According to this report, 10 publications had an EID. But at least 32 additional publications in our Pure publication set from December had an EID as no other UID is assigned to them in our Pure raw data and they were matched in SciVal.

The unmatched publications with a UID are shown in Table 4b. PMID is the most common UID, up to 90%. Close to 40% of the publications have a DOI. For the December publication set, we assume that the unmatched publications have no EID, otherwise they should have been matched.

In the following three tables we analyzed publication type as registered in Pure. Notable, but not surprising, is that close to 100% of

Table 6a. Language of publications with UID exported from Pure to SciVal and matched. Download from August and December 2015.

	August 1770 publications		December 1844 publications	
English	1766	>99%	1817	99%
Other	4	<1%	27	1%

Table 6b. Language of publications with UID exported from Pure to SciVal and not matched. Download from August and December 2015.

	August 244 publications		December 178 publications	
English	127	52%	77	43%
Other	117	48%	101	57%

Table 6c. Language of publications without UID not exported from Pure to SciVal. Download from August and December 2015.

	August 54 publications		December 44 publications	
English	26	48%	23	52%
Other	28	52%	21	48%

matched publications are journal contributions (Table 5a), as these are usually well-represented in Scopus/SciVal. What is surprising, however, is that practically the same percentage of unmatched publications with a UID is journal contributions (Table 5b). For the publications without a UID (Table 5c) there are still many journal publications, approximately 60%, but a much lower share than for the publications with a UID. The distributions among publication types do not differ substantially between the August and December publication sets. All publication sets include very few non-journal publications.

We also analyzed the languages of the publications. Concerning the matched publications, 99% are written in English. Interestingly, the absolute number of matched publications in other languages increased from 4 to 27 between August and December (Table 6a). Elsevier's automatic assignment of EIDs may improve the match for these publications in our specific setting. However, our publication set is far too small to draw any conclusions from. For the unmatched publications with and without UID in the August and December publication sets, the ratios between English and other languages are close to fifty-fifty (Table 6b and Table 6c).

Our analysis reveals more journal publications in English not matched in SciVal than we expected. Therefore we extracted lists of the top journals according to number of publications from our

Table 7a. Top journals according to number of publications, for publications with UIDs exported from Pure to SciVal and matched. Download from August and December 2015.

August 1770 publications		December 1844 publications	
Journal title	**Number of publications**	**Journal title**	**Number of publications**
PLOS ONE	62	*PLOS ONE*	65
Contact Dermatitis	27	*Danish Medical Journal*	38
Danish Medical Journal	27	*Contact Dermatitis*	27
BMJ Open	17	*BMJ Open*	19

Table 7b. Top journals according to number of publications for publications with UIDs exported from Pure to SciVal and not matched. Download from August and December 2015.

August 244 publications		December 178 publications	
Journal title	**Number of publications**	**Journal title**	**Number of publications**
Ugeskrift for Læger, Ugeskrift for Laeger	114	*Ugeskrift for Læger, Ugeskrift for Laeger*	99
Danish Medical Journal	10	*Clinical and Translational Allergy*	3
Cochrane Database of Systematic Reviews	7	*PLOS ONE*	3
PLOS ONE	6	*Annals of Clinical and Translational Neurology, EJNMMI Physics, Endocrine Connections, and Oncoimmunology* (all 2 publications each)	2

Table 7c. Top journals according to number of publications for publications without UIDs not exported from Pure to SciVal. Download from August and December 2015.

August 54 publications		December 44 publications	
Journal title	**Number of publications**	**Journal title**	**Number of publications**
Ugeskrift for Læger, Ugeskrift for Laeger	14	*Ugeskrift for Læger, Ugeskrift for Laeger*	8
Clinical Nutrition	3	*Clinical Nutrition*	3
Early Intervention in Psychiatry	3	*Early Intervention in Psychiatry*	3
Journal of Anesthesia & Clinical Research	2	*American Journal of Nuclear Medicine and Molecular Imaging, Annals of Internal Medicine, Annals of Sports Medicine and Research, Bibliotek for Laeger, European Respiratory Journal, International Journal of Anatomy and Research, Journal of Anesthesia & Clinical Research, Journal of Clinical Toxicology, Journal of Gastroenterology and Hepatology Research, Klinisk Sygepleje, Læring og Medier - LOM* (all 1 publication each)	1

publication sets. For the unmatched publications a large share is published in the two journals of the Danish Medical Association (*Ugeskrift for Læger* and *Danish Medical Journal*). Both are indexed by Scopus. Interestingly, we see *Ugeskrift for Læger* represented among the matched publications, the unmatched publications with UID, and the unmatched publications without UID. Also *PLOS ONE* publications are among both the matched and the unmatched publications, but not among publications without a UID. Three of six unmatched *PLOS ONE* publications from the August publication set are matched in December. The remaining three *PLOS ONE* publications were still not registered in Scopus in December. Table 7b and Table 7c includes more journals which are indexed by Scopus but the publications are not matched. For example *Clinical Nutrition* (cf. Table 7c) with 299 publications from 2014 indexed in Scopus, and *Clinical and Translational Allergy* (cf. Table 7b) with only 4 publications from 2014 indexed in Scopus. This may indicate some shortcomings in the Scopus indexing procedures. For our publication sets this seemed to be the biggest problem for a successful matching of publications with an UID. These results suggest that the missed matches were due to missing publications in Scopus. Another explanation could be missing UIDs in Scopus. This was detected in a study from 2012 (Bar-Ilan *et al.*, 2012, p. 101). In addition to this, 35 publications from the August and December publication sets were from journals not indexed by Scopus according to Scopus' Content Coverage Guide.

In summary, the literature review shows that only a few studies report findings on UIDs as match keys. Results on coverage are reported and errors in the matching procedure are less frequently addressed (Table 1).

The findings from the case study show that the majority of the publications were matched (85.6% in August and 89.3% in December). Almost all the matched publications have a DOI and are journal publications in English. Among the matched publications, 67 (3.8%) in the publication set from August have a duplicate DOI, whereas 32 (1.7%) from December do. Other error types (Table 1) were observed which lowered the precision of the match between Pure and SciVal. Still, duplicate DOIs are the most prevalent problem. However, both coverage and precision have improved from August to December. This can be explained to some extent by Scopus's automatic merging of Article in Press and the published version. Elsevier's procedure of adding EIDs to publications in Pure may correct other duplicates and improve the coverage. Finally, duplicates may have been corrected manually by Elsevier in Scopus.

The unmatched publications also include journal publications. Close to half of these are in Danish and published in the journal *Ugeskrift for Læger* of which the indexing in Scopus is highly irregular. Our analysis indicates that journals with publications in English also suffer from similar irregular indexing but to a much lesser extent.

The matching of publications across databases can be used to create publication sets for bibliometric analysis. As an example we now briefly discuss how publication sets formed by using UIDs as match keys may affect the bibliometric indicators number of publications, number of citations, and the average number of citations

per publication. This is to our knowledge only discussed briefly in the two studies. They both conclude that duplicate DOIs can lead to errors in bibliometric analysis (Franceschini *et al.*, 2015, p. 2186; Valderrama-Zurián *et al.*, 2015, p. 575).

The coverage can affect bibliometric indicators. Results from our case study indicated that the majority of the publications from Pure are matched correctly in SciVal. Yet, the difference between the August and the December publication sets and the analysis of top journals (Table 7a–Table 7c) show that coverage can be improved. This means that the number of publications and citations could be higher in a bibliometric analysis based on our publication set. We do not know the number of citations for the publications not matched but based on our knowledge about journals not indexed fully by Scopus we discuss the potential consequences for number of citations. *Ugeskrift for Læger* has over 100 publications that are not covered in Scopus/SciVal. The journal is not highly cited (Scopus 2014 IPP = 0.127, SNIP = 0.109) so inclusion of the missing publications would probably increase the number of citations a little, but lower the average number of citations per publication. However, inclusion of the missing publications for other journals could potentially have the opposite effect and increase the average number of citations per publication. An example is *PLOS ONE* (Scopus 2014 IPP = 3.270, SNIP = 1.034).

The precision of a bibliometric indicator is distorted by the fact that some DOIs are matched multiple times in SciVal. In most cases it is due to a duplicate of the same publication, but we also observed instances of publications in our Pure publication set with a DOI duplicate in SciVal not present in the Pure set (Table 2). The duplicates have several implications for the bibliometric indicators number of publications, number of citations, and the average number of citations per publication:

The number of publications becomes inflated by inclusion of duplicates. In our publication sets from August and December the publication count increased by 3.8% and 1.7%, respectively. Therefore we recommend that when the number of publications is calculated, duplicates should be removed whether the duplicate publication is in the original Pure publication set or not.

Before citations are counted, all duplicates not present in the Pure publication set must be deleted. For the remaining duplicate pairs we found that sometimes both duplicates were cited independently. In all instances except one there was no overlap between the citations. Citations divided between duplicates in Scopus are also reported in another study where variations of a journal name results in duplicates in Scopus. It is suggested that databases like Scopus can improve verification of DOIs to solve the duplicate problem (Valderrama-Zurián *et al.*, 2015). Publications were from 2014 and the export from Pure to SciVal was done in August and December 2015. So the publications had a very short time to attract citations. In total the publication set from August had 4,982 citations, but if citations for the duplicates were included the number was 5,047. This is an additional 1.3% citations. For the publication set from December the total number of citations has increased to 7,695, and to 7,720 if citations for duplicates are included. The duplicates account for 0.3% additional citations. In our study, many of the duplicates removed were Article in Press whereas the Article version was kept.

The calculation of the average number of citations per publication should not include duplicates in counting publications but include duplicates in counting citations. If duplicates are kept the average number of citations per publication will be too low. If the duplicates are removed some of the citations may also be discarded and again the average number of citations per publication will be too low.

Dataset 1. Data of evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal.

http://dx.doi.org/10.5256/f1000research.8913.d126923

Data sets consisting of publications from the Department of Clinical Medicine (DoCM) at University of Copenhagen (UCPH) are provided. A description of each file is provided in 'Dataset description'.

Conclusion

UIDs are seen as effective keys to match identical publications across databases or identify duplicates in a database. The use of UIDs as match keys is well-implemented in many settings but only few studies evaluate how UIDs work as match keys. As DOIs are implemented in more and more settings DOI also becomes increasingly interesting as a match key. According to the publication years of the studies in our literature review we suggest that this trend took off around 2010.

Our case study confirms the findings of the literature review. UIDs as match keys do not return a 100% coverage of a publication set, and include errors for a small part of the matches. It is not possible to draw conclusions on when the coverage and precision is satisfactory as this should be discussed in relation to the purpose of a matching exercise, exemplified here as a bibliometric analysis.

In our case study we identified irregular indexing of journals in Scopus as the main problem for an optimal coverage. Duplicate DOIs were a particular problem for the precision of UIDs as match keys. This type of error is easy to detect while other types of errors demand a more thorough analysis of the publication sets. This analysis could be done by using a traditional match key based on title, author name, etc. Other error types also present in our case study are: incorrect UIDs in reference lists and databases, and UIDs not registered by the database where a bibliometric analysis is performed.

Match keys to detect identical publications across databases are used for many purposes, but our focus is bibliometric indicators. Here the duplicate DOIs constitute a problem as both keeping them in the publication set to improve the reliability of citation counts and deleting them to improve the reliability of publication counts will distort the calculation of average number of citations per publication and the many other bibliometric indicators which combine publication and citation counts. Also the coverage of a

publication set can affect bibliometric indicators. We have discussed that failing to fully cover a low or high impact journal may also lead to imprecise bibliometric indicators.

Future implications

Our purpose has been to contribute to the discussion on how well UIDs work as match keys for publications with a focus on preparing publication sets for bibliometric analysis. Compared to traditional match keys where bibliographic information is used, UIDs are efficient, but they also have drawbacks.

The coverage of UIDs is fully dependent on whether a UID is assigned to a publication, and its representations in publication lists and databases. Here the traditional match key has an advantage as it often is dependent on basic bibliographic data and can be modified to fit different formats. The traditional match key will probably have a good chance of retrieving all publications with a UID if the representations of the publications have basic bibliographic data of a fair quality. In addition, the traditional match key can retrieve publications without UIDs.

The precision of UIDs depends on how carefully a UID is assigned to a publication and its representations in publication lists and databases. Using a single UID as a match key can be fragile as no crosschecks are made on other data fields. Detection of errors requires an examination of the result of the matching process. The traditional match key often relies on more data fields and thus has a built-in crosscheck. Neither of the match keys will solve the problem of duplicates of identical publications.

We recommend more studies to be done on the pros and cons of UIDs because UIDs are being increasingly introduced for more entities and adopted as efficient match keys. The availability of UIDs may become critical for the inclusion of a publication or a database in a bibliometric analysis.

Data availability

F1000Research: Dataset 1. Data of evaluation of unique identifiers used as keys to match identical publications in Pure and SciVal. 10.5256/f1000research.8913.d126923 (Gauffriau *et al.*, 2016).

Author contributions

The authors have contributed to the work according to the four criteria for authorship in *Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals.* MG designed the work and performed the literature review. HHM and DM performed the experiments. All authors analyzed the data, drew conclusions, wrote and edited the paper. All authors agree on the final content.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

The authors wish to thank Guillaume Warnan and Floortje Flippo from Elsevier's SciVal team for information on procedures in Pure, Scopus, and SciVal as well as manual matching of publications between Pure and SciVal. We are also grateful for the chance to present and discuss a preliminary version of our case study at the 20th Nordic Workshop on Bibliometrics and Research Policy 2015 in Oslo. And last but not least, thank you to Lorna Wildgaard from the Royal School of Library and Information Science, Denmark, for commenting on an earlier version of this paper. And thank you to the reviewers invited by F1000Research for their valuable comments which have resulted in a second and improved version of the paper.

References

Altmetric.com. (n.d.): **Embeddable badges.** Retrieved August 11, 2015.
Reference Source

arXiv.org: **Understanding the arXiv identifier.** 2015.
Reference Source

Costas R, Zahedi Z, Wouters P: **Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective.** *J Assoc Inf Sci Technol.* 2015; **66**(10): 2003–2019.
Publisher Full Text

CrossRef: **cited-by linking.** Retrieved August 9, 2015. 2015a.
Reference Source

CrossRef: **Publishers and societies.** Retrieved December 6, 2015, 2015b.
Reference Source

DataCite. (n.d.): **About DataCite.** Retrieved February 13, 2016.
Reference Source

Elsevier: **Scopus comes of age.** Retrieved December 6, 2015, 2004.
Reference Source

Elsevier: **Elsevier Enhances Pure, Providing New Research Analysis Functionalities through Direct Integration with SciVal.** Retrieved August 8, 2015. 2014.
Reference Source

Franceschini F, Maisano D, Mastrogiacomo L: **A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics.** *J Am Soc Inf Sci Technol.* 2013; **64**(10): 2149–2156.
Publisher Full Text

Franceschini F, Maisano D, Mastrogiacomo L: **Scientific journal publishers and omitted citations in bibliometric databases: Any relationship?** *J Informetr.* 2014; **8**(3): 751–765.
Publisher Full Text

Franceschini F, Maisano D, Mastrogiacomo L: **Errors in DOI indexing by bibliometric databases.** *Scientometrics.* 2015; **102**(3): 2181–2186.
Publisher Full Text

Gauffriau M, Madsen HH, Madsen D: **Dataset 1 in: Evaluation of unique**

identifiers used for citation linking. *F1000Research.* 2016.
Data Source

Hammerton JA, Granitzer M, Harvey D, *et al.*: **On generating large-scale ground truth datasets for the deduplication of bibliographic records.** In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics - WIMS '12.* New York, USA: ACM Press, 2012; 18.
Publisher Full Text

Haunschild R, Bornmann L: **Normalization of Mendeley reader counts for impact assessment.** *J Informetr.* 2016; **10**(1): 62–73.
Publisher Full Text

Haustein S, Bowman TD, Macaluso B, *et al.*: **Measuring Twitter activity of arXiv e-prints and published papers.** In *altmetrics14: expanding impacts and metrics.* 2014.
Publisher Full Text

Haustein S, Siebenlist T: **Applying social bookmarking data to evaluate journal usage.** *J Informetr.* 2011; **5**(3): 446–457.
Publisher Full Text

HEFCE: **The Metric Tide: Correlation analysis of REF2014 scores and metrics.** 2015.
Publisher Full Text

International DOI Foundation: **DOI® Handbook.** Retrieved August 11, 2015, 2015.
Reference Source

Jiang Y, Lin C, Meng W, *et al.*: **Rule-based deduplication of article records from bibliographic databases.** *Database (Oxford).* 2014; **2014**(0): bat086.
PubMed Abstract | Publisher Full Text | Free Full Text

Kim KY, Kim HM: **A Study on Developing and Refining a Large Citation Service System.** *International Journal of Knowledge Content Development & Technology.* 2013; **3**(1): 65–80.
Publisher Full Text

Kraker P, Enkhbayar A, Lex E: **Exploring Coverage and Distribution of Identifiers on the Scholarly Web.** 2015.
Reference Source

Moed HF, Debruin RE, Vanleeuwen TN: **New Bibliometric Tools for the Assessment of National Research Performance - Database Description, Overview of Indicators and First Applications.** *Scientometrics.* 1995; **33**(3): 381–422.
Publisher Full Text

Mongeon P, Paul-Hus A: **The journal coverage of Web of Science and Scopus: a comparative analysis.** *Scientometrics.* 2016; **106**(1): 213–228.
Publisher Full Text

Nuredini K, Peters I: **Economic and Business Studies Journals and Readership Information from Mendeley.** In F. Pehar, C. Schloegl, & C. Wolff (Eds.), *Re:inventing Information Science in the Networked Society: Proceedings of the 14th International Symposium on Information Science.* Zadar. 2015.
Reference Source

Olensky M: **Data accuracy in bibliometric data sources and its impact on citation matching.** Humboldt-Universität zu Berlin. 2014.
Reference Source

Olensky M, Schmidt M, van Eck NJ: **Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science.** *J Assoc Inf Sci Technol.* 2015.
Publisher Full Text

Ostell JM, Wheelan SJ, Kans JA: **The NCBI Data Model.** In A. D. Baxevanis & B. F. F. Ouellette (Eds.), *Bioinformatics.* (second Edi.,). New York, USA: John Wiley & Sons, Inc. 1998; 19–43.
Publisher Full Text

Research Councils UK: **Interoperability Pilot Phase 1 Report Purpose of the Pilot.** 2015.
Reference Source

Schneider J, Bloch CW, Aagaard K, *et al.*: **Analyses of the scholarly and scientific output from grants funded by the Danish Council for Independent Research from 2005–2008.** 2014.
Reference Source

Valderrama-Zurián JC, Aguilar-Moya R, Melero-Fuentes D, *et al.*: **A systematic analysis of duplicate records in *Scopus*.** *J Informetr.* 2015; **9**(3): 570–576.
Publisher Full Text

Wilsdon J, Allen L, Belfiore E, *et al.*: **The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management.** 2015.
Publisher Full Text

Zahedi Z, Costas R, Wouters P: **How well developed are altmetrics? A cross-disciplinary analysis of the presence of "alternative metrics" in scientific publications.** *Scientometrics.* 2014a; **101**(2): 1491–1513.
Publisher Full Text

Zahedi Z, Haustein S, Bowman TD: **Exploring data quality and retrieval strategies for Mendeley reader counts.** In *ASIS&T Workshop on Informetric and Scientometric Research: Metrics14.* 2014b.
Reference Source

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 2

Referee Report 23 September 2016

doi:10.5256/f1000research.10259.r16140

✓ **Keith G. Jeffery**
Keith G. Jeffery Consultants, Faringdon, UK

The authors have addressed the criticisms made adequately (and also those of the other referee) and so I approve of the article being indexed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 02 September 2016

doi:10.5256/f1000research.9591.r15520

✓ **Sarah L. Shreeves**
University of Miami, Coral Gables, FL, USA

This article describes a small study to better understand how unique identifiers (in this case DOIs) assigned to publications work as a match key for what the authors call 'citation linking' or matching identical publications in different databases in order to form a publication set that can then be used for bibliometric analysis. The study's goals were to lay out the characteristics of publications that were matched versus those that were unmatched, and to illustrate how the publications set formed might be affected by problems in the matching.

- I found the discussion of the citation linking throughout the article confusing, in part because my expectation of what that means is quite different than what is described in the article. My understanding of citation linking is that one is following works cited in a publication via the UID. It took several reads to clarify this; it may be something that the authors wish to clarify further in the abstract or introduction.

- I found the study design, methods, and conclusions to be adequate for the study particularly for the first objective (given the acknowledgement that this is not generalizable). It would be useful to expand this study to other disciplines to see whether there are similar issues and characteristics. In

order to replicate this study in another discipline, however, more attention may need to be paid to coverage in Pure and SciVal, or use of other databases may need to be deployed.

- The second objective- to better understand how bibliometric analysis of the publication set formed might be skewed by issues in matching- was undermined by the size of the dataset- there seemed to be too little data from which to draw firm conclusions. However, given that, the description of the results are fine.

- I believe that the Future Implications section provides the clearest implications of the study - essentially that UIDs aren't panaceas for deduplication/ citation matching/ etc.- and that use of UIDs as match points should be informed by potential points of failure. This description could have clarified other pieces of the article.

I wavered between 'Approved' and 'Approved with Reservations', and settled on 'Approved'. I believe that the revisions necessary are really clarity of language and noting throughout that the size and scope of the case study provide a limited window.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 19 Sep 2016

Marianne Gauffriau, University of Copenhagen

Thank you for your review. You touch upon important points. Please see our response below.

Sarah L. Shreeves: “I found the discussion of the citation linking throughout the article confusing, in part because my expectation of what that means is quite different than what is described in the article. My understanding of citation linking is that one is following works cited in a publication via the UID. It took several reads to clarify this; it may be something that the authors wish to clarify further in the abstract or introduction.”

The term *citation linking*: In the literature we saw an inconsistent use of terms, but settled on "citation linking" to mean the matching of representations of identical publications across databases. We got similar comments from other reviewers and have edited our article, minimizing the term "citation linking", instead using "match key" and "matching process".

Sarah L. Shreeves: “I found the study design, methods, and conclusions to be adequate for the study particularly for the first objective (given the acknowledgement that this is not generalizable). It would be useful to expand this study to other disciplines to see whether there are similar issues and characteristics. In order to replicate this study in another discipline, however, more attention may need to be paid to coverage in Pure and SciVal, or use of other databases may need to be deployed.”

Generalization: We have focused on Pure and SciVal in this study and plan to do the same in a follow-up study where we expand the data sets to cover more research fields within health sciences and investigate a new matching process introduced by Elsevier. We did not consider

including other databases as the main focus in the present study is the matching process. Thus we have selected a case where we expected the coverage to be good. If other disciplines with a low coverage in SciVal were studied, we agree that it would be important to analyze and discuss the coverage in more detail. As pointed out in your review, additional databases could be considered.

Sarah L. Shreeves: “The second objective- to better understand how bibliometric analysis of the publication set formed might be skewed by issues in matching- was undermined by the size of the dataset- there seemed to be too little data from which to draw firm conclusions. However, given that, the description of the results are fine.“

Bibliometric indicators and the discussion of them are based on a limited data set. We agree: Our results apply to our case alone and may serve as inspiration for other similar studies. In version 2 of the article the part on bibliometric indicators is elaborated upon. See the subsection *Results of the case study* in the paper, version 2.

Sarah L. Shreeves: “I believe that the Future Implications section provides the clearest implications of the study - essentially that UIDs aren't panaceas for deduplication/ citation matching/ etc.- and that use of UIDs as match points should be informed by potential points of failure. This description could have clarified other pieces of the article.”

Future Implications: In the section *Future Implications* we discuss UIDs as match key compared to a traditional match key. We hope to address this in a follow-up study, following Elsevier changing the way publications are matched between (Scopus and) Pure and SciVal: Publications in Pure are matched with publications in Scopus using traditional match keys and assigned a Scopus ID of the matched publication. The match between Pure and SciVal now happens using Scopus ID as only match key.

Sarah L. Shreeves: “I wavered between 'Approved' and 'Approved with Reservations', and settled on 'Approved'. I believe that the revisions necessary are really clarity of language and noting throughout that the size and scope of the case study provide a limited window.”

Again thank you for your review. We hope that our response and the edited paper have addressed your comments.

Competing Interests: No competing interests were disclosed.

Referee Report 08 August 2016

doi:10.5256/f1000research.9591.r15523

? **Keith G. Jeffery**
Keith G. Jeffery Consultants, Faringdon, UK

This paper aims to contribute to the discussion on use of UIDs for citation linking. It states two objectives:

1. Explore the coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match key.

2. Illustrate how publication sets formed by using UIDs as the match key may affect the bibliometric indicators: Number of publications, number of citations and the average number of citations per publication.

The paper usefully lists the common UIDs used for research publications, and in so doing illustrates the problem of disjoint UID sets (intersected by other attributes of the publication) and lack of universal uniqueness but without commenting upon it. An important facility in PURE (based on CERIF) is the ability to utilise 'federated IDs' so that there can be several 'unique identifiers' for the same object thus allowing crosswalking within the metadata for a given publication. The paper mentions UIDs for publications and persons but not for organisations; the use of such UIDs is becoming more common and is useful for comparative analysis of performance by research organisational units.

The paper states (P3) "Opposite to the other UIDs, the DOI is assigned to the publication itself and not merely to the publication's representation in a database". There is a lack of clarity of thinking here. The DOI is a digital identifier and thus is not assigned to the publication itself (the scholarly content) but to a digital representation (textual, diagrammatic, tabular...) of it. The representation in the database is of two parts: the metadata (traditionally the library catalog entry) and the publication content (possibly with added datasets, software or artifacts). Throughout the paper it should be made clear that the work is based on the metadata - although the citation is within the textual content of the publication referring to a reference at the end of the publication from where a link can/should be made to the full text of that (cited) article. In fact ideally the link should have a time period of validity and semantics of the kind of citation (such as negative or positive, explanatory...).

The paper states (P4) "The matching of identical publications in different databases is called citation linking or reference linking. Matching within the same database is called deduplication. The term citation matching is also used, but often for the more specific purposes where citing and cited publications are matched". It is important that the paper clarifies the thinking here also. Citation has nothing to do with matching or de-duplication, although 'clean' data is required for effective and efficient citation. In fact the paper does not really discuss citation itself but only the utilisation of UIDs to achieve citation - and more particularly the difficulties of obtaining 'clean' data.

The paper does not reference "Citation Linking: Improving Access to Online Journals"
S. Hitchcock*, L. Carr, S. Harris, J. M. N. Hey and W. Hall *Proceedings of the 2nd ACM International Conference on Digital Libraries*, edited by Robert B. Allen and Edie Rasmussen, 1997 (New York, USA: Association for Computing Machinery), pp. 115-122 available at
http://journals.ecs.soton.ac.uk/acmdl97.htm which states clearly (and as early as 1997):

"The likely effect of citation linking can be gauged by recognising a direct parallel with citation indexing, which is sometimes referred to as 'forward' referencing (i.e. for a given article, all the subsequent papers that cite it), developed by Garfield (1955). Garfield's description of citation indexing as an 'association of ideas' bears remarkable similarity to Bush's 'association of thoughts' which anticipated modern hypertext. Citation linking combines the two approaches, mapping both reference data and citation index data on to the text in the form of links. Adding electronic links to the literature is introducing a new culture in many respects, but citation linking is likely to be acceptable to the academic community because it builds on practices established in other forms such as print. It also allows the community to exploit its own intellectual input in this process, recognised in Garfield's original rationale that "by using authors' references in compiling the citation index, we are in reality utilizing an army of indexers"." Here citation

linking is used - and with precedence - to mean something quite different to the use in this paper.

The literature review is appropriate. The case study is limited in organisational unit, research discipline, time period; however the results are useful and could - with advantage - be compared with similar studies in other disciplines /organisational units and time periods.

Overall the paper is a useful contribution and hopefully will stimulate similar studies in other disciplines and time ranges. but the lack of clarity of thinking should be addressed. I would suggest:

1. a paragraph clarifying the inter-relationships of metadata and content for one publication with that for another cited from the first;

2. a paragraph explaining citation, citation linking, citation indexing;

3. the existing paragraph on UIDs being expanded to explain the usefulness of 'federated IDs' for crosswalking within the metadata of a given publication.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 25 Aug 2016

Marianne Gauffriau, University of Copenhagen

Response to Marion Schmidt and Keith G. Jeffery

Thank you for the valuable reviews. Below we quote the issues raised in the reviews followed by our response and description of how we have edited the paper. Please note that we have sent the same reply to both of you so you have the same information on how we have edited the paper.

Marion Schmidt: “In the introduction and literature review, it should be made clearer that the authors use a specific definition of the term citation linking (linking items between databases) for their study and it should be clarified if other studies refer to the same or other scenarios (like reference matching or deduplication).”

Keith G. Jeffery: “The paper states (P4) "The matching of identical publications in different databases is called citation linking or reference linking. Matching within the same database is called deduplication. The term citation matching is also used, but often for the more specific purposes where citing and cited publications are matched". It is important that the paper clarifies the thinking here also. Citation has nothing to do with matching or de-duplication, although 'clean' data is required for effective and efficient citation. In fact the paper does not really discuss citation itself but only the utilisation of UIDs to achieve citation - and more particularly the difficulties of obtaining 'clean' data.
The paper does not reference "Citation Linking: Improving Access to Online Journals" S. Hitchcock*, L. Carr, S. Harris, J. M. N. Hey and W. Hall Proceedings of the 2nd ACM International Conference on Digital Libraries, edited by Robert B. Allen and Edie Rasmussen, 1997 (New York, USA: Association for Computing Machinery), pp. 115-122 available at http://journals.ecs.soton.ac.uk/acmdl97.htm which states clearly (and as early as 1997):

"The likely effect of citation linking can be gauged by recognising a direct parallel with citation indexing, which is sometimes referred to as 'forward' referencing (i.e. for a given article, all the subsequent papers that cite it), developed by Garfield (1955). Garfield's description of citation indexing as an 'association of ideas' bears remarkable similarity to Bush's 'association of thoughts' which anticipated modern hypertext. Citation linking combines the two approaches, mapping both reference data and citation index data on to the text in the form of links. Adding electronic links to the literature is introducing a new culture in many respects, but citation linking is likely to be acceptable to the academic community because it builds on practices established in other forms such as print. It also allows the community to exploit its own intellectual input in this process, recognised in Garfield's original rationale that "by using authors' references in compiling the citation index, we are in reality utilizing an army of indexers"." Here citation linking is used - and with precedence - to mean something quite different to the use in this paper.
[…]
I would suggest:
[…]
2. a paragraph explaining citation, citation linking, citation indexing;"

The term *citation linking* has caused confusion as citation can be understood as the reference from one paper (citing document) to another paper (cited document). Also in the literature we see an inconsistent use of the term. By citation linking we have meant the linking of representations of identical publications across databases. We have had difficulties finding an unambiguous term for this and ended with citation linking. We now have minimized the use of the term citation linking. Instead we use match key and matching process.

Marion Schmidt: "The sample is a convenience sample based only on one year and one subject field (health sciences), with a predominantly journal- and English-based publication culture. The authors describe their first objective as "Explore the coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match key", after introducing the more general goal to investigate "how well UIDs work for citation linking" (p.3). In order to truly answer the goals and objectives, the hypothesis should be discussed and tested that implementation of UIDs may vary across disciplines, depending on the amount of e.g. smaller, regional journals and easy data import options for a system like Pure from, e.g. PubMed. In this perspective a careful sampling strategy representative for all subjects that are, to a varying extent, covered in the target database SciVal/Scopus as well as more publication years, would have been much more fruitful in order to assess the coverage and thus usability of UIDs in both systems. Thus, data from a university-wide implementation of Pure would make up a reasonable case study.
The authors deal inconsistently with this issue as they try to rectify their sample on the one hand "As this type of publication and the health sciences are well-covered in Scopus/SciVal [...], we expected the DoCM publication set to be well-fitted for our purpose, namely to explore the citation linking process, rather than how well SciVal covers publications from a department" (p.4) - which is a non-adequate argument as the coverage could be considered separately. On the other hand, they concede that their "case study" may not lead to generalizable results and that results will therefore be compared to those from a literature review (p.4).
[…]
I would strongly suggest expanding the initial sample to other fields."

Keith G. Jeffery: "The case study is limited in organisational unit, research discipline, time period; however the results are useful and could - with advantage - be compared with similar studies in other disciplines /organisational units and time periods."

The case we work with was deliberately chosen as the best scenario and not as a sample fit for generalization. We wanted to study how well UIDs work as match key in the integration between Pure and SciVal, so it seemed beneficial to work with a sample with as many UIDs registered as possible. Because of the attributes of publications from the Department of Clinical Medicine (journal publications in English), they seemed likely to have a high number of UIDs registered. Furthermore, the integration between Pure and SciVal is relatively new and we only had limited knowledge about how it worked and found no evaluations of the integration in the literature. We know that publication sets from health sciences are well-covered by Scopus. With the potential good coverage between Pure and SciVal we hoped that our publication set would allow us primarily to focus on the matching itself. The coverage of SciVal/Scopus is reported in the paper, but this is supporting information.
In the selection of the case we also took local conditions into account. Working with Pure data validated by the University Library, that is publications from 2014, we felt confident that the sample would have fewer errors regarding registration/correct registration of UIDs than samples from previous years where no central validation routines were in place.
It is not possible for us to repeat our analysis with a case covering more research areas and more publication years, as the matching process for the integration between Pure and SciVal has changed. Now only Scopus ID is used as match key. Before, DOI and PubMed ID also worked as match keys. However, we are doing a follow-up study to see how the new matching process works for our original publication set and a new larger publication set.

Marion Schmidt: “Thus, title and abstract information do - in my opinion - not reflect the actual limits of the study adequately.”

The objective was formulated somewhat broadly and could give the impression that the study included a thorough analysis of the coverage of SciVal/Scopus and the study yielded results which could be generalized to other
research fields. As explained above, this is not the case. We have specified the objective to clearly show that we evaluate how well UIDs work as match key. Our focus is the matching process and any errors in this process. Furthermore, we do only work with one case – the integration between Pure and SciVal for a publication set from health sciences. The case cannot be generalized to other research fields or settings. As a consequence of the reformulated objective, we have changed the title and abstract accordingly.

Keith G. Jeffery: “The paper usefully lists the common UIDs used for research publications, and in so doing illustrates the problem of disjoint UID sets (intersected by other attributes of the publication) and lack of universal uniqueness but without commenting upon it. An important facility in PURE (based on CERIF) is the ability to utilise 'federated IDs' so that there can be several 'unique identifiers' for the same object thus allowing crosswalking within the metadata for a given publication. The paper mentions UIDs for publications and persons but not for organisations; the use of such UIDs is becoming more common and is useful for comparative analysis of performance by research organisational units.
The paper states (P3) "Opposite to the other UIDs, the DOI is assigned to the publication itself and not merely to the publication’s representation in a database". There is a lack of clarity of thinking here. The DOI is a digital identifier and thus is not assigned to the publication itself (the scholarly content) but to a digital representation (textual, diagrammatic, tabular...) of it. The representation in the database is of two parts: the metadata (traditionally the library catalog entry) and the publication content (possibly with added datasets, software or artifacts). Throughout the paper it

should be made clear that the work is based on the metadata - although the citation is within the textual content of the publication referring to a reference at the end of the publication from where a link can/should be made to the full text of that (cited) article. In fact ideally the link should have a time period of validity and semantics of the kind of citation (such as negative or positive, explanatory...).
[…]
I would suggest:
1. a paragraph clarifying the inter-relationships of metadata and content for one publication with that for another cited from the first;
[…]
3. the existing paragraph on UIDs being expanded to explain the usefulness of 'federated IDs' for crosswalking within the metadata of a given publication."

The description of a publication and its UIDs is elaborated. We did not analyze the full content of a publication but only parts of its metadata representations in Pure and SciVal. A publication can have more UIDs assigned to it in one or both of the databases. We assumed that each of the UIDs for a publication in a database represented the same publication. See *Introduction* and the subsection *Limitations* in the paper, version 2.

Marion Schmidt: "On the other hand, they concede that their "case study" may not lead to generalizable results and that results will therefore be compared to those from a literature review (p.4). This claim, however, is not really fulfilled, as the studies mentioned in the literature review use UIDs for different purposes, constellations and databases. More importantly, no real comparison takes place, as the authors only recap "but all conclude that UIDs do not cover all records in the databases" (p.6). In order to contextualize their own results, concrete settings and results of other studies should be represented and discussed.
[…]
In the introduction and literature review, it should be made clearer that the authors use a specific definition of the term citation linking (linking items between databases) for their study and it should be clarified if other studies refer
to the same or other scenarios (like reference matching or deduplication)."

Keith G. Jeffery: "The literature review is appropriate."

The literature review shows the big picture and now also includes more details on the studies included. None of the studies in the review were directly comparable to our study as we analyzed how well UIDs work as match key in the matching process between Pure and SciVal. We now show that the studies in the review used UIDs for many different matching processes. Only two of the studies focused on the evaluation of UIDs. Still, based on all the studies we summarized information on error types when UIDs are used as match keys, which is relevant for our study. See the subsection *Results of the literature review* in the paper, version 2.

Marion Schmidt: "With regard to the second purpose "Illustrate how publication sets formed by using UIDs as the match key may affect the bibliometric indicators: Number of publications, number of citations and the average number of citations per publication", the authors do not actually calculate citation indicators including and excluding publications (which are covered, but could not be matched via UIDs), but discuss the problem mostly theoretically."

The number of citations was only calculated for the publications matched in SciVal at the time

we did the exports from Pure to SciVal (August and December 2015). We did not have the number of citations for the other publications. This information is added to the article. For the matched publications we have added the number of citations including and excluding duplicates. See the subsection *Results of the case study* in the paper, version 2.

Marion Schmidt: “Methods and data (with exceptions mentioned separately below) are sufficiently clearly documented, but the whole study lacks generalizability and, partly, elaborateness of analyses, as in case of the citation indicator perspective. Besides, given the rather manageable amounts of unmatched publications in this study, a comprehensive and more elaborate search and analysis of the causes of the missed match (not covered, missing UID in Scopus, ..) would be preferable. In larger corpora, this could be done via a random sample.”

Causes for missed matches for publications with a UID were not investigated when we did the exports from Pure to SciVal at publication level but only at journal level (Table 7a-7c). We have looked into the question now, but many of the publications not matched in August and December 2015 are now matched. We added to the paper that from our analysis of journals (Table 7a-7c) it is likely that the publications not matched were not indexed in Scopus. We identified three *PLOS ONE* publications which were not indexed in Scopus, and *Clinical and Translational Allergy* had only four publications from 2014 indexed in Scopus. But the missed matches can also be due to missing UIDs in Scopus. For 35 publications we established that they were published in journals not indexed in Scopus. This is also added to the paper. See the subsection *Results of the case study* in the paper, version 2.

Marion Schmidt: “The authors write "In the integration between Pure and SciVal DOI, PMID and EID are used as match keys for citation linking. From March 2016 the integration between Pure and SciVal is based on EID alone. This will not affect the present study as we analyze publication sets downloaded in August and December 2015" (p.3).
The authors should discuss what this does mean with respect of the relevance of their results. Could it mean that the matching of current publications will be probably better?”

The new matching process between Pure and SciVal means that our study is a documentation of the first matching processes between Pure and SciVal and can serve as background for new evaluations. We have added this point to the paper. We do not know if the new functionality is better than the ones we have evaluated but we are looking into that in a follow-up study. See the *Introduction* in the paper, version 2

Marion Schmidt: “Regarding the publication type categorization: Has the categorization been informed by some available classification or, e.g. database scheme? In my opinion, some mappings are sub-optimal and particular; especially the assignment of Doctoral Thesis to "Other" instead of "Book", Encyclopedia chapter to "Other" instead of "Contribution to Book", Editorial to "Other" instead of "Journal", Article in Proceedings to "Journal" instead of "Contribution to Book" (please compare the WoS document type and publication type classification for the third and fourth case). Including so many and different publication types in a residual category is unprofitable for the later analysis.”

The categorization of publication types has been reorganized according to the categories in Pure. See Table 5a-c.

Marion Schmidt: “I did not understand the fact that EIDs seem to have been attributed

automatically without being visible in the raw data. How?"

The EID automatically attributed to publications in Pure was introduced with Pure version 4.23.0. EIDs are added to publications in Pure as an automatic job. In the beginning the job ran every day with a sub-set of the publications in order not to overload the matching service. 90 days after installation of the new version of Pure the job had covered all publications. Now when a new publications are added to Pure they will be included in the first coming match, hence within 14 days.
The automatic attributed EIDs are not shown in our raw data (excel files downloaded from Pure). Therefore it is not possible to detect which publications are matched on the automatic attributed EIDs. More publications are matched in the November dataset, and we assume they are matched on the automatic attributed EID but cannot prove it.

Once again thank you for the reviews. We hope that our response and the edited paper have addressed all your comments.

Competing Interests: No competing interests were disclosed.

Referee Report 28 July 2016

doi:10.5256/f1000research.9591.r14675

? **Marion Schmidt**
German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

The sample is a convenience sample based only on one year and one subject field (health sciences), with a predominantly journal- and English-based publication culture. The authors describe their first objective as "Explore the coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match key", after introducing the more general goal to investigate "how well UIDs work for citation linking" (p.3). In order to truly answer the goals and objectives, the hypothesis should be discussed and tested that implementation of UIDs may vary across disciplines, depending on the amount of e.g. smaller, regional journals and easy data import options for a system like Pure from, e.g. PubMed. In this perspective a careful sampling strategy representative for all subjects that are, to a varying extent, covered in the target database SciVal/Scopus as well as more publication years, would have been much more fruitful in order to assess the coverage and thus usability of UIDs in both systems. Thus, data from a university-wide implementation of Pure would make up a reasonable case study.

The authors deal inconsistently with this issue as they try to rectify their sample on the one hand "As this type of publication and the health sciences are well-covered in Scopus/SciVal [...], we expected the DoCM publication set to be well-fitted for our purpose, namely to explore the citation linking process, rather than how well SciVal covers publications from a department" (p.4) - which is a non-adequate argument as the coverage could be considered separately. On the other hand, they concede that their "case study" may not lead to generalizable results and that results will therefore be compared to those from a literature review (p.4). This claim, however, is not really fulfilled, as the studies mentioned in the literature review use UIDs for different purposes, constellations and databases. More importantly, no real comparison takes place, as the authors only recap "but all conclude that UIDs do not cover all records in the databases" (p.6). In order to contextualize their own results, concrete settings and results of other studies should be represented and discussed.

With regard to the second purpose "Illustrate how publication sets formed by using UIDs as the match key may affect the bibliometric indicators: Number of publications, number of citations and the average number of citations per publication", the authors do not actually calculate citation indicators including and excluding publications (which are covered, but could not be matched via UIDs), but discuss the problem mostly theoretically.

I would strongly suggest expanding the initial sample to other fields.

Thus, title and abstract information do - in my opinion - not reflect the actual limits of the study adequately.

Methods and data (with exceptions mentioned separately below) are sufficiently clearly documented, but the whole study lacks generalizability and, partly, elaborateness of analyses, as in case of the citation indicator perspective. Besides, given the rather manageable amounts of unmatched publications in this study, a comprehensive and more elaborate search and analysis of the causes of the missed match (not covered, missing UID in Scopus, ..) would be preferable. In larger corpora, this could be done via a random sample.

In the introduction and literature review, it should be made clearer that the authors use a specific definition of the term citation linking (linking items between databases) for their study and it should be clarified if other studies refer to the same or other scenarios (like reference matching or deduplication).

The authors write "In the integration between Pure and SciVal DOI, PMID and EID are used as match keys for citation linking. From March 2016 the integration between Pure and SciVal is based on EID alone. This will not affect the present study as we analyze publication sets downloaded in August and December 2015" (p.3).

The authors should discuss what this does mean with respect of the relevance of their results. Could it mean that the matching of current publications will be probably better?

Regarding the publication type categorization: Has the categorization been informed by some available classification or, e.g. database scheme? In my opinion, some mappings are sub-optimal and particular; especially the assignment of Doctoral Thesis to "Other" instead of "Book", Encyclopedia chapter to "Other" instead of "Contribution to Book", Editorial to "Other" instead of "Journal", Article in Proceedings to "Journal" instead of "Contribution to Book" (please compare the WoS document type and publication type classification for the third and fourth case). Including so many and different publication types in a residual category is unprofitable for the later analysis.

I did not understand the fact that EIDs seem to have been attributed automatically without being visible in the raw data. How?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 25 Aug 2016

Marianne Gauffriau, University of Copenhagen

Response to Marion Schmidt and Keith G. Jeffery

Thank you for the valuable reviews. Below we quote the issues raised in the reviews followed by our response and description of how we have edited the paper. Please note that we have sent the same reply to both of you so you have the same information on how we have edited the paper.

Marion Schmidt: “In the introduction and literature review, it should be made clearer that the authors use a specific definition of the term citation linking (linking items between databases) for their study and it should be clarified if other studies refer to the same or other scenarios (like reference matching or deduplication).”

Keith G. Jeffery: “The paper states (P4) "The matching of identical publications in different databases is called citation linking or reference linking. Matching within the same database is called deduplication. The term citation matching is also used, but often for the more specific purposes where citing and cited publications are matched". It is important that the paper clarifies the thinking here also. Citation has nothing to do with matching or de-duplication, although 'clean' data is required for effective and efficient citation. In fact the paper does not really discuss citation itself but only the utilisation of UIDs to achieve citation - and more particularly the difficulties of obtaining 'clean' data.
The paper does not reference "Citation Linking: Improving Access to Online Journals" S. Hitchcock*, L. Carr, S. Harris, J. M. N. Hey and W. Hall Proceedings of the 2nd ACM International Conference on Digital Libraries, edited by Robert B. Allen and Edie Rasmussen, 1997 (New York, USA: Association for Computing Machinery), pp. 115-122 available at http://journals.ecs.soton.ac.uk/acmdl97.htm which states clearly (and as early as 1997):
"The likely effect of citation linking can be gauged by recognising a direct parallel with citation indexing, which is sometimes referred to as 'forward' referencing (i.e. for a given article, all the subsequent papers that cite it), developed by Garfield (1955). Garfield's description of citation indexing as an 'association of ideas' bears remarkable similarity to Bush's 'association of thoughts' which anticipated modern hypertext. Citation linking combines the two approaches, mapping both reference data and citation index data on to the text in the form of links. Adding electronic links to the literature is introducing a new culture in many respects, but citation linking is likely to be acceptable to the academic community because it builds on practices established in other forms such as print. It also allows the community to exploit its own intellectual input in this process, recognised in Garfield's original rationale that "by using authors' references in compiling the citation index, we are in reality utilizing an army of indexers"." Here citation linking is used - and with precedence - to mean something quite different to the use in this paper.
[…]
I would suggest:
[…]
2. a paragraph explaining citation, citation linking, citation indexing;”

The term *citation linking* has caused confusion as citation can be understood as the reference from one paper (citing document) to another paper (cited document). Also in the literature we see an inconsistent use of the term. By citation linking we have meant the linking of representations of identical publications across databases. We have had difficulties finding an unambiguous term for this and ended with citation linking. We now have minimized the use of the term citation linking. Instead we use match key and matching process.

Marion Schmidt: “The sample is a convenience sample based only on one year and one subject field (health sciences), with a predominantly journal- and English-based publication culture. The authors describe their first objective as "Explore the coverage, precision, and characteristics of publications matched versus not matched with UIDs as the match key", after introducing the more general goal to investigate "how well UIDs work for citation linking" (p.3). In order to truly answer the goals and objectives, the hypothesis should be discussed and tested that implementation of UIDs may vary across disciplines, depending on the amount of e.g. smaller, regional journals and easy data import options for a system like Pure from, e.g. PubMed. In this perspective a careful sampling strategy representative for all subjects that are, to a varying extent, covered in the target database SciVal/Scopus as well as more publication years, would have been much more fruitful in order to assess the coverage and thus usability of UIDs in both systems. Thus, data from a university-wide implementation of Pure would make up a reasonable case study.
The authors deal inconsistently with this issue as they try to rectify their sample on the one hand "As this type of publication and the health sciences are well-covered in Scopus/SciVal [...], we expected the DoCM publication set to be well-fitted for our purpose, namely to explore the citation linking process, rather than how well SciVal covers publications from a department" (p.4) - which is a non-adequate argument as the coverage could be considered separately. On the other hand, they concede that their "case study" may not lead to generalizable results and that results will therefore be compared to those from a literature review (p.4).
[…]
I would strongly suggest expanding the initial sample to other fields.”

Keith G. Jeffery: “The case study is limited in organisational unit, research discipline, time period; however the results are useful and could - with advantage - be compared with similar studies in other disciplines /organisational units and time periods.”

The case we work with was deliberately chosen as the best scenario and not as a sample fit for generalization. We wanted to study how well UIDs work as match key in the integration between Pure and SciVal, so it seemed beneficial to work with a sample with as many UIDs registered as possible. Because of the attributes of publications from the Department of Clinical Medicine (journal publications in English), they seemed likely to have a high number of UIDs registered. Furthermore, the integration between Pure and SciVal is relatively new and we only had limited knowledge about how it worked and found no evaluations of the integration in the literature. We know that publication sets from health sciences are well-covered by Scopus. With the potential good coverage between Pure and SciVal we hoped that our publication set would allow us primarily to focus on the matching itself. The coverage of SciVal/Scopus is reported in the paper, but this is supporting information.
In the selection of the case we also took local conditions into account. Working with Pure data validated by the University Library, that is publications from 2014, we felt confident that the sample would have fewer errors regarding registration/correct registration of UIDs than samples from previous years where no central validation routines were in place.
It is not possible for us to repeat our analysis with a case covering more research areas and more publication years, as the matching process for the integration between Pure and SciVal has changed. Now only Scopus ID is used as match key. Before, DOI and PubMed ID also worked as match keys. However, we are doing a follow-up study to see how the new matching process works for our original publication set and a new larger publication set.

Marion Schmidt: “Thus, title and abstract information do - in my opinion - not reflect the actual

limits of the study adequately."

The objective was formulated somewhat broadly and could give the impression that the study included a thorough analysis of the coverage of SciVal/Scopus and the study yielded results which could be generalized to other research fields. As explained above, this is not the case. We have specified the objective to clearly show that we evaluate how well UIDs work as match key. Our focus is the matching process and any errors in this process. Furthermore, we do only work with one case – the integration between Pure and SciVal for a publication set from health sciences. The case cannot be generalized to other research fields or settings. As a consequence of the reformulated objective, we have changed the title and abstract accordingly.

Keith G. Jeffery: "The paper usefully lists the common UIDs used for research publications, and in so doing illustrates the problem of disjoint UID sets (intersected by other attributes of the publication) and lack of universal uniqueness but without commenting upon it. An important facility in PURE (based on CERIF) is the ability to utilise 'federated IDs' so that there can be several 'unique identifiers' for the same object thus allowing crosswalking within the metadata for a given publication. The paper mentions UIDs for publications and persons but not for organisations; the use of such UIDs is becoming more common and is useful for comparative analysis of performance by research organisational units.
The paper states (P3) "Opposite to the other UIDs, the DOI is assigned to the publication itself and not merely to the publication's representation in a database". There is a lack of clarity of thinking here. The DOI is a digital identifier and thus is not assigned to the publication itself (the scholarly content) but to a digital representation (textual, diagrammatic, tabular...) of it. The representation in the database is of two parts: the metadata (traditionally the library catalog entry) and the publication content (possibly with added datasets, software or artifacts). Throughout the paper it should be made clear that the work is based on the metadata - although the citation is within the textual content of the publication referring to a reference at the end of the publication from where a link can/should be made to the full text of that (cited) article. In fact ideally the link should have a time period of validity and semantics of the kind of citation (such as negative or positive, explanatory...).
[…]
I would suggest:
1. a paragraph clarifying the inter-relationships of metadata and content for one publication with that for another cited from the first;
[…]
3. the existing paragraph on UIDs being expanded to explain the usefulness of 'federated IDs' for crosswalking within the metadata of a given publication."

The description of a publication and its UIDs is elaborated. We did not analyze the full content of a publication but only parts of its metadata representations in Pure and SciVal. A publication can have more UIDs assigned to it in one or both of the databases. We assumed that each of the UIDs for a publication in a database represented the same publication. See *Introduction* and the subsection *Limitations* in the paper, version 2.

Marion Schmidt: "On the other hand, they concede that their "case study" may not lead to generalizable results and that results will therefore be compared to those from a literature review (p.4). This claim, however, is not really fulfilled, as the studies mentioned in the literature review use UIDs for different purposes, constellations and databases. More importantly, no real comparison takes place, as the authors only recap "but all conclude that UIDs do not cover all

records in the databases" (p.6). In order to contextualize their own results, concrete settings and results of other studies should be represented and discussed.
[…]
In the introduction and literature review, it should be made clearer that the authors use a specific definition of the term citation linking (linking items between databases) for their study and it should be clarified if other studies refer to the same or other scenarios (like reference matching or deduplication)."

Keith G. Jeffery: "The literature review is appropriate."

The literature review shows the big picture and now also includes more details on the studies included. None of the studies in the review were directly comparable to our study as we analyzed how well UIDs work as match key in the matching process between Pure and SciVal. We now show that the studies in the review used UIDs for many different matching processes. Only two of the studies focused on the evaluation of UIDs. Still, based on all the studies we summarized information on error types when UIDs are used as match keys, which is relevant for our study. See the subsection *Results of the literature review* in the paper, version 2.

Marion Schmidt: "With regard to the second purpose "Illustrate how publication sets formed by using UIDs as the match key may affect the bibliometric indicators: Number of publications, number of citations and the average number of citations per publication", the authors do not actually calculate citation indicators including and excluding publications (which are covered, but could not be matched via UIDs), but discuss the problem mostly theoretically."

The number of citations was only calculated for the publications matched in SciVal at the time we did the exports from Pure to SciVal (August and December 2015). We did not have the number of citations for the other publications. This information is added to the article. For the matched publications we have added the number of citations including and excluding duplicates. See the subsection *Results of the case study* in the paper, version 2.

Marion Schmidt: "Methods and data (with exceptions mentioned separately below) are sufficiently clearly documented, but the whole study lacks generalizability and, partly, elaborateness of analyses, as in case of the citation indicator perspective. Besides, given the rather manageable amounts of unmatched publications in this study, a comprehensive and more elaborate search and analysis of the causes of the missed match (not covered, missing UID in Scopus, ..) would be preferable. In larger corpora, this could be done via a random sample."

Causes for missed matches for publications with a UID were not investigated when we did the exports from Pure to SciVal at publication level but only at journal level (Table 7a-7c). We have looked into the question now, but many of the publications not matched in August and December 2015 are now matched. We added to the paper that from our analysis of journals (Table 7a-7c) it is likely that the publications not matched were not indexed in Scopus. We identified three *PLOS ONE* publications which were not indexed in Scopus, and *Clinical and Translational Allergy* had only four publications from 2014 indexed in Scopus. But the missed matches can also be due to missing UIDs in Scopus. For 35 publications we established that they were published in journals not indexed in Scopus. This is also added to the paper. See the subsection *Results of the case study* in the paper, version 2.

Marion Schmidt: "The authors write "In the integration between Pure and SciVal DOI, PMID and

EID are used as match keys for citation linking. From March 2016 the integration between Pure and SciVal is based on EID alone. This will not affect the present study as we analyze publication sets downloaded in August and December 2015" (p.3).
The authors should discuss what this does mean with respect of the relevance of their results. Could it mean that the matching of current publications will be probably better?"

The new matching process between Pure and SciVal means that our study is a documentation of the first matching processes between Pure and SciVal and can serve as background for new evaluations. We have added this point to the paper. We do not know if the new functionality is better than the ones we have evaluated but we are looking into that in a follow-up study. See the *Introduction* in the paper, version 2

Marion Schmidt: "Regarding the publication type categorization: Has the categorization been informed by some available classification or, e.g. database scheme? In my opinion, some mappings are sub-optimal and particular; especially the assignment of Doctoral Thesis to "Other" instead of "Book", Encyclopedia chapter to "Other" instead of "Contribution to Book", Editorial to "Other" instead of "Journal", Article in Proceedings to "Journal" instead of "Contribution to Book" (please compare the WoS document type and publication type classification for the third and fourth case). Including so many and different publication types in a residual category is unprofitable for the later analysis."

The categorization of publication types has been reorganized according to the categories in Pure. See Table 5a-c.

Marion Schmidt: "I did not understand the fact that EIDs seem to have been attributed automatically without being visible in the raw data. How?"

The EID automatically attributed to publications in Pure was introduced with Pure version 4.23.0. EIDs are added to publications in Pure as an automatic job. In the beginning the job ran every day with a sub-set of the publications in order not to overload the matching service. 90 days after installation of the new version of Pure the job had covered all publications. Now when a new publications are added to Pure they will be included in the first coming match, hence within 14 days.
The automatic attributed EIDs are not shown in our raw data (excel files downloaded from Pure). Therefore it is not possible to detect which publications are matched on the automatic attributed EIDs. More publications are matched in the November dataset, and we assume they are matched on the automatic attributed EID but cannot prove it.

Once again thank you for the reviews. We hope that our response and the edited paper have addressed all your comments.

Competing Interests: No competing interests were disclosed.